

Claims

1. A system for annotating sets and subsets of genes, comprising:

- (a) means for identifying a set of genes;
- (b) means for partitioning the set of genes in (a) into subsets known as clusters;
- (c) means for associating a set of documents with each gene in the set of genes in (a), and consequently a means for associating a set of documents with each of the clusters in (b);
- (d) means for receiving the text of part or all of each of the documents in (c);
- (e) means for assigning numerical weights to words or phrases contained in the text in (d), said assignment being made by partitioning the documents according to their association with each cluster as provided in (b) and (c), followed by the application, to the words or phrases in those partitioned documents, of any of the word weight-setting methods that are implemented in the computer program Rainbow;
- (f) means for sorting, storing, and displaying the words and phrases contained in documents associated with each of the clusters provided in (b), said sorting being based on the numerical weights assigned to the words as provided in (e), and said storage and display allowing words and phrases to be arranged in the order of their sorted numerical weights;

whereby the words and phrases having the greatest numerical weights for each cluster provide an indication of the concepts, structures, functions, and processes with which said cluster is most particularly associated.

2. A system for evaluating the quality of gene clustering, comprising:
 - (a) means for identifying a set of genes;
 - (b) means for partitioning the set of genes in (a) into subsets known as clusters;
 - (c) means for associating a set of documents with each gene in the set of genes in (a), and consequently a means for associating a set of documents with each of the clusters in (b);
 - (d) means for partitioning the set of documents in (c) into two subsets, a training subset and a testing subset;
 - (e) means for receiving the text of part or all of each of the training subset documents in (d);
 - (f) means for receiving the text of part or all of each of the testing subset documents in (d);
 - (g) means for using words or phrases in the text of documents in (e) to train a document classifier, said training being accomplished by partitioning the documents according to their association with each cluster as provided in (b) and (c), followed by the parameter-fitting, using the words or phrases in those partitioned documents, of any of the document classifiers that are implemented in the computer program Rainbow;
 - (h) means for using words or phrases in the text of each document in (f) to test the trained document classifier in (g), wherein the classifier predicts the cluster with which the test document is associated;
 - (i) means for the option of calculating and storing the fractions of test documents in (d) known to correspond to each cluster as provided in (b) and (c), that are correctly

predicted to be associated with each cluster, upon testing with the document classifier as provided in (h);

- (j) means for the option of repeatedly and randomly partitioning documents in (c) into training and test subsets as provided in (d), for using each such partitioning to calculate a fraction of correct classifications for each cluster as provided in (e)-(i), and for storing said fractions for each and every such random partitioning of documents into training and test subsets.
- (k) means for the option of repeatedly and randomly partitioning the set of genes in (a) into subsets, wherein the sizes of the random subsets are matched to the sizes of the clusters as provided in (b); for re-associating a set of documents with each gene in the set as in (c), and consequently associating a set of documents with each of the randomly partitioned subsets of genes; for making available means (d)-(i) so as to be able to calculate a fraction of correct classifications for each of the random partitions that are matched to the clusters as provided in (b); and for storing said fractions for each and every such random partitioning of the set of genes in (a).
- (l) means for the option of calculating a measure of central tendency, such as mean or median, for the fractions that were generated by repeated, random partitioning of documents in (j), and for the fractions that were generated by repeated, random partitioning of the set of genes in (k); and for calculating a figure-of-merit for each cluster as the numerical difference between said measure of central tendency obtained from (j) and (k);

whereby said figure-of-merit for each cluster provides an indication of the extent to which some words and phrases, present in documents associated with genes in said cluster, collectively distinguish that cluster from all the other clusters, and whereby said figure-of-merit for each cluster provides an indication of the extent to which the annotations produced by the system of Claim 1 distinguish the clusters, and whereby said figure-of-merit for each cluster provides an indication of the quality of that cluster.

3. A system for evaluating the quality of gene clustering, comprising:

- (a) means for identifying a set of genes;
- (b) means for partitioning the set of genes in (a) into subsets known as clusters;
- (c) means for associating a set of documents with each gene in the set of genes in (a);
- (d) means for calculating for every pair of genes within a cluster in (b) a coupling strength index, said index being proportional to the number of times that any document in (c) is associated with both members of said pair of genes; and for storing said set of index values for every cluster;
- (e) means for repeatedly and randomly partitioning the set of genes in (a) into subsets, wherein the sizes of the random subsets are matched to the sizes of the clusters as provided in (b); for re-associating a set of documents with each gene in the set as in (c); for making available means (d) so as to be able to calculate coupling strength indices; and for storing said set of index values for every such random subset;
- (f) means for calculating for every cluster in (b) and random subset in (e) a measure of

time-varying functions at each of the measurement time points for each

one of said plurality of genes; and

wherein f and k are approximated as truncated Taylor series, the coefficients of which

are estimated using the received data x , at each of the measurement time points

for each one of said plurality of genes; and

wherein the estimated values of the synthesis rate, f , at each of the measurement

time points for each one of said plurality of genes are used to cluster said plurality

of genes;

whereby said clustering may reveal subsets of genes among the plurality of genes

that are regulated by the same transcription factors, as evidenced by the similarity

of their time-varying transcription rates.

5. The method of claim 4, used as the means for partitioning a set of genes into clusters in a system for annotating sets and subsets of genes.

6. The method of claim 4, used as the means for partitioning a set of genes into clusters in a system for evaluating the quality of gene clustering.